Volume 4, Issue 7, July 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Devanagri Handwritten Text Character Segmentation Techniques and Related Issues - A Review

Er. Binny Thakral
Student, M.Tech. (CE)
Yadavindra College of Engineering,
Talwandi Sabo, Punjab, India

Er. Manoj Chaudhary
Assistant Professor, CE Dept.,
Yadavindra College of Engineering,
Talwandi Sabo, Punjab, India

Abstract: Segmentation is a crucial part of Optical Character Recognition (OCR). OCR is methodology of changing the scanned images of Machine printed or Manually written contents, images, letters into format which might be spoken to or prepared by computer as ASCII. OCR could be utilized for automated processing and handling of forms, old corrupted reports, bank cheques, postal codes and structures. Segmentation is method that partitions the printed or handwritten content into individual lines, words or characters. The primary test in segmentation of handwritten information is that there is wide mixed bag of varieties of styles and handwritings. The principle point of this review paper is to illuminate the diverse features of Devanagri script, the existing character segmentation techniques and the issues in handwritten text that makes segmentation more troublesome.

Keywords: OCR, Character Segmentation, Devanagri script, Offline, Touching characters.

I. INTRODUCTION

Segmentation is one of the testing and challenging fields in OCR. It is an operation that tries to divide an image of arrangement of characters into sub images of individual symbols. It is one of the decision procedures in OCR. In text report image examination, the significant step is extraction of text lines from documents, and afterward the text lines are divided into words and characters.

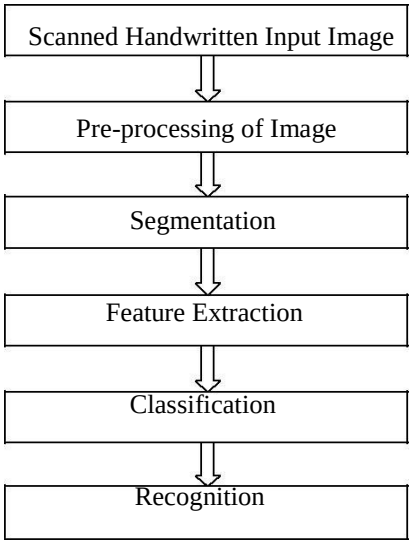

Fig. 1: Optical Character Recognition System

Categorization of OCR is focused around two primary criteria: Data Acquisition process and Type of Text composed. Data Acquisition process incorporates online and offline information, where the online information means handwriting that is recorded with a digitizer, as a period gathering of pen headings and offline speaks to the hard duplicate of handwriting scanned by optical scanner or camera. Text Type incorporates Machine printed and Handwritten content, Machine printed content holds the materials, for example, books, daily papers, magazines, archives, and different composition units in the feature or still image. Machine printed characters are uniform in stature, width, and pitch means the same textual style and size are utilized. Handwritten incorporates the content composed by distinctive writers by their hands. There are loads of varieties in handwriting of diverse users.

II. CHARACTERISITICS OF DEVANAGRI SCRIPT

There are numerous scripts and languages in India yet not much research has been carried out towards recognition of Handwritten Indian characters. Devanagri Script is utilized within Hindi, Nepali, Marathi, Sindhi, Sanskrit and Konkani languages. Hindi is most prominent dialect in India and third most popular language in world, so much research work is possible on Hindi language recognition. No upper and lower case idea is there in Hindi as in English dialect. It is phonetic and syllabic script, the words are composed in Hindi as they are proclaimed and syllabic content is composed by consonants and vowels that together structure syllables.

There are 33 Consonants, 13 vowels and 14 modifiers in Hindi, an expression could be vertically separated into three parts: the Upper area, the middle area and the lower area as shown in figure 2:

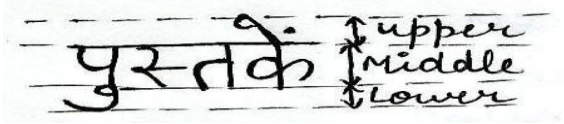

Fig 2: Three regions of Devanagri word

Middle region contains vowels, consonants or combination of both and the upper & lower regions contain vowel symbols or parts of modifiers.

Vowels could be composed as independent or by utilizing various diacritical marks, which are composed above, underneath, before or after the consonant or vowel they fit in with, these are called as Modifiers or 'Matras'. Frequently two or more consonants combine to structure another shape this is called as compound character. Blend of half and full consonants can structure 'Conjuct' Characters. In every conjunct character, the right part is a full consonant, and the left part is dependably a half consonant. One unique feature of Devanagri is flat horizontal line on top of all characters, called Header line or 'Shirorekha'. Header line of one character joins the header line of past and next character in order to structure a full word.

III. EXSITING APPROACHES FOR CHARACTER SEGMENTATION

Presence of lots of irregularities in handwritten Devanagri text like skewed, broken, touching and overlapping data causes various difficulties for segmenting characters efficiently. Various existing approaches and their respective outcomes are summarized below:

3.1 Character segmentation is considered as an important area of OCR process. Richard *et al*. 1996 [1] in their paper discussed the various methods and strategies used in character segmentation. Segmentation methods are divided into four main categories: Classical Approach, Recognition based, Holistic approach and Hybrid method. They have discussed various dissection techniques, projection analysis methods, connected component processing methods and various recognition based techniques.

3.2 Satish *et al*. 2010 [2] gave the details about discrepancies and problematic areas in successful character segmentation. Author widely discussed about devanagri script and the areas that causes the problems in recognition process of handwritten data. Bad handwriting and lack of language knowledge leads to various irregularities that makes recognition of hand written text more challenging. These problems includes: inaccurate use of modifiers in upper and lower region, their abnormal size, incomplete representation and wrong insertion of header line.

3.3 A new technique for segmentation of Devanagri documents using Histogram Approach is proposed in Vikas *et al.* 2011 [3]. Preprocessing before segmentation is performed on the input images by Thresholding, Noise Reduction, Skew Correction and Thinning methods. Horizontal, Vertical projection and Histograms for the images are used to calculate the white pixels in each column. Then find the position of single white pixel and replace these columns by 1, after inverting the image and making the columns as 0, mark the bounding box for characters and copy the pixels in bounding boxes in different file. These different files contain the segmented characters. This approach results in 55% of accuracy of character segmentation but it needs more effort and segmentation of compound characters is still to be carried out.

3.4 The variations in writing styles of different users make handwritten character segmentation more difficult task. Ashwin *et al.* 2012 [4], proposed new algorithm for segmentation of lines, words and characters. Preprocessing includes smoothing of and binarization of image. Line segmentation is carried out by connected component approach in which connected components are clustered to extract line. Then vertical projection profile is used to segment words by dividing the spaces. Base characters are segmented by vertical profile using clear paths within them. This algorithm works well for non-overlapping and unbroken characters with 97% of accuracy.

3.5 Kunal *et al.* 2013 [5], proposed new approach for segmentation of Devanagri Characters using Neighbourhood Tracing Algorithm and Projection Profiles. Authors introduced two passes for algorithm, in which pass 1includes scanning of whole document so as to extract an individual word and the scanning is performed from top to bottom and from left to right. After detecting first black pixel, it is assumed that word has been detected and then by tracing the boundary, that is extracted from the text. Then pass 2 removes shirorekha from that word and Projection Profile techniques are used to segment the characters from the word. This algorithm can only used be for segmenting simple isolated characters in words, a lot of work can be done to segment fused and touching characters.

3.6 Touching characters in a word make the character segmentation more challenging task. Shuchi *et al.* 2014 [6], proposed the technique for fragmentation of handwritten touching characters by detecting joint points and forming bounding boxes over each character. Joint points are the meeting point's means where two characters meet. Characters are separated using these joint points to find vertical bars and header lines. To fragment each character in a word

bounding boxes are formed. Within each word horizontal rectangles are identified. In thinned image bounding boxes are formed after removing the joint points. Each word is categorized in three parts: Mid bar, Side bar and Non bar. In Mid Bar, if width of bounding box is greater than single character width, then make a cut at leftmost joint point of large bounding box and separate the characters. In Side Bar, when two side bar characters are touching each other, form bounding box. If there is no large and no equal size bounding boxes, then make a cut line after each detected vertical bars. In Non Bar, if character without Vertical Box then make cut at the right of Bounding Box. The average result of this technique is 71%; still a lot of efficiency is needed.

3.7 Munish *et al*. 2014 [7] describes the character segmentation of offline handwritten data of touching characters using water reservoir method. Preprocessing phase includes Skew Detection/Correction, Skeltonization and Noise removal Steps. The given Algorithm uses Horizontal and Vertical Projection Techniques for identification of touching characters. The horizontal and vertical projection profiles techniques together with water reservoir technique have been applied on all documents, in which the reservoirs means the cavity areas where water i.e. one of ones can be stored as reservoir and these reservoirs specifies the touching characters . Vertical projection is used for finding and removing the header line, so as to extract the sub images properly. The touching Characters are segmented with 93.51% of accuracy, still efficiency needs to be improved and this technique can also be applicable on Devanagri touching characters in future.

IV. ISSUE ZONES IN CHARACTER SEGMENTATION

There are different irregularities in handwritten segmentation which are performed by author in composing that makes the content hard to recognize. Because of hurried or awful handwriting, absence of focus and ill-advised learning about language, terrible environment and bad stationary can result in segmentation process all the more difficult. Major discrepancy areas are as follows:

a) Strange size of upper and lower modifiers: Sometimes modifiers or "matras" are not written as indicated by the best possible size just because of the varieties in composing styles of different writers. Such kind of irregularities in modifiers causes issues in recognition of character or word.

Fig 3: Variations in writing same Modifier

b) Incomplete characters: Incomplete and inaccurate written words or characters by people reasons real challenges in recognition for readers and additionally machine.

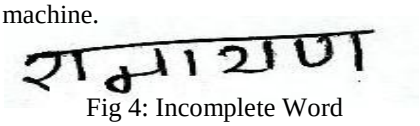

Fig 4: Incomplete Word

c) Mixing of upper/lower modifiers with center area: When the modifiers of upper and lower region blends with the characters of center region, a monstrous representation turns out, and it gets hard to recognize for a machine to understand that in which area it ought to be set.

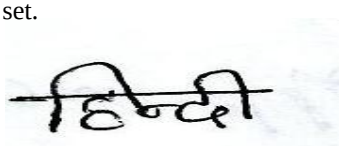

Fig 5: Upper Modifier mixing with middle region Character

d) Unsymmetrical header line or Shirorekha: Devanagri characters are joined with each other utilizing header line. A tilted, slanted or rushed header line makes critical issues for recognition framework.

Fig 6: Unsymmetrical Header line

e) Skewed words: Sometimes words are not composed in straight way, they may be slanted upward or descending, this reasons issues in identifying the header line of that character.

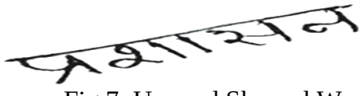

Fig 7: Upward Skewed Word

f) Broken characters: A few segments of the characters may be missing which represents to a solitary character as more than one character. The characters may be broken up in horizontal or vertical form as:

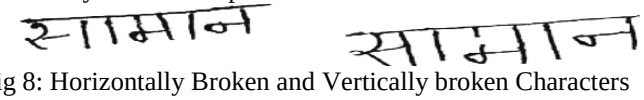

Fig 8: Horizontally Broken and Vertically broken Characters

g) Overlapped characters: Due the awful handwriting now and again the characters in a word cover one another, means a segment of one character or modifier is intermixed with the an alternate character, which makes issue in recognition. An example of overlapping characters is shown in fig. 9:

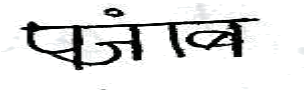

Fig 9: Second character overlapping previous one

h) Touching words: Some individuals while writing touch the characters with one another, by this machine couldn't understand the character is single or joined to other one, which again causes trouble in recognition. These might be separated into: Intra-touching characters and Inter-touching characters.

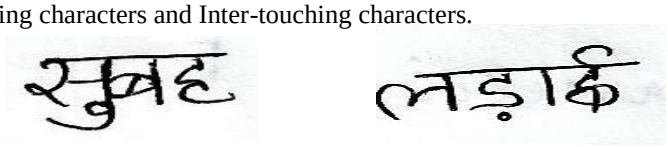

Fig 10: Inter-touching and Intra-touching Characters

i) Overwritten characters: Sometimes an author overwrites the piece of character or stroke as to guarantee that character ought not to be broken. Machine takes it as more broaden character and these reasons issues.

Fig 11: Overwritten Character and Modifier

j) Slanted characters: A few people are having inclined composition style, they write the content in tilted way such that the header line goes from lowest part to up rather that in straight level way. So this makes trouble in perceiving the header line.

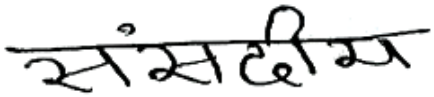

Fig 12: Slanted Characters

V. CONCLUSION

Presence of lots of irregularities in handwritten information makes handwritten Character Segmentation and recognition more troublesome as compared with printed information. This causes more issues on OCR process. Different problematic zones have been illuminated in this paper. A great deal of research work is possible on these issue areas. OCR process might be enhanced by instructing individuals to overcome hasty handwriting and also time to time researchers need to enhance their algorithms for significant improvement in character segmentation.

REFERENCES

[1] Richard G. Casey, Eric Lecolinet, "A Survey of Methods and Strategies in Character Segmentation", IEEE Transactions on Patterm Analysis and Machine Intelligence, Vol. 18, No.17, July 1996.

[2] Satish Kumar, "An Analysis of Irregularities in Devanagri Script Writing—A Machine Recognition Perspective", International Journal on Computer Science and Engineering, Vol. 2, No. 2, PP: 274-279, 2010.

[3] Vikas J. Dongre, Vijay H. Mankar, "Devanagri Document Segmentation Using Histogram approach", International Journal of Computer Science, Engineering and Information Technology, Vol. 1, No. 3, PP: 46-53, Aug, 2011.

[4] Ashwin S Ramteke, Milind E Rane, "Offline Handwritten Devanagari Script Segmentation", International Journal of Scientific & Technology Research, Vol. 1, Issue 4, PP: 142-145, May, 2012.

[5] Kunal Shah, Jaideep Singh, Prashant Pushkarna, Hasnain Kurawadwala, Abhishek Alate, "A New Approach for Segmentation of Devanagri Characters", Vol. 2, Issue 4, PP: 162-164, 2013.

[6] Shuchi Kapoor, Vivek Verma, "Fragmentation of Handwritten Touching Characters in Devanagri Script", International Journal of Information Technology, Modeling and Computing (IJITMC) Vol. 2, No. 1, PP: 11-21, Feb 2014.

[7] Munish Kumar, M.K. Jindal, R.K. Sharma, "Segmentation of Isolated and Touching Characters in Offline Handwritten Gurmukhi Script Recognition", International Journal Information Technology and Computer Science, PP: 58-63, Feb, 2014.

[8] Naresh Kumar Garg, Lakhwinder Kaur, M.K. Jindal, "Segmentation of Handwritten Hindi Text", International Journal of Computer Applications, Vol. 1, No. 4, PP: 19-22, 2010.

[9] Veena Bansal, R.M.K. Sinha, "Segmentation of Touching Characters in Devanagri", Technical Report, TDIL,

IIT Kanpur, India.

[10] M. K. Jindal, G. S. Lehal, R. K. Sharma, "A Study of Touching Characters in Degraded Gurmukhi Text", Proceedings of World Academy of Science, Engineering and Technology, Vol. 4, Feb, 2005.

[11] Rajiv Kumar, Amardeep Singh, "Detection and Segmentation of Lines and Words in Gurmukhi Handwritten Text",
2nd International Advance Computing Conference, IEEE 2010.

[12] Sharma, Palaiahnakote Shivakumara, Umapada Pal, Michael Blumenstein and Chew Lim Tan, "A New Method for
Character Segmentation from Multi-Oriented Video Words", 12th International Conference on Document Analysis and Recognition, IEEE, 2013.

[13] U. Pal, Sagarika Datta, "Segmentation of Bangla Unconstrained Handwritten Text", Proceedings of the Seventh International Conference on Document Analysis and Recognition, IEEE 2003.

[14] Galaxy Bansal, Dharamveer Sharma, "Isolated Handwritten Words Segmentation Techniques in Gurmukhi Script",
International Journal of Computer Applications (0975 - 8887) Volume 1 – No. 24, PP: 104-111, 2010.

[15] Dipak K. Koshti, Sharvari Govilkar, "Segmentation of Touching Characters in Handwritten Devanagri Script", International Journal of Computer Science and its Applications, Vol. 2, Issue 2, PP: 83-87.

